Classification of Newborn EEG Maturity with Bayesian Averaging over Decision Trees

V. Schetinin, L. Jakaite
Department of Computer Science and Technology
University of Bedfordshire, Luton, LU1 3JU, UK
Vitaly.Schetinin@beds.ac.uk, +4407770918445

Abstract – EEG experts can assess a newborn's brain maturity by visual analysis of age-related patterns in sleep EEG. It is highly desirable to make the results of assessment most accurate and reliable. However, the expert analysis is limited in capability to provide the estimate of uncertainty in assessments. Bayesian inference has been shown providing the most accurate estimates of uncertainty by using Markov Chain Monte Carlo (MCMC) integration over the posterior distribution. The use of MCMC enables to approximate the desired distribution by sampling the areas of interests in which the density of distribution is high. In practice, the posterior distribution can be multimodal, and so that the existing MCMC techniques cannot provide the proportional sampling from the areas of interest. The lack of prior information makes MCMC integration more difficult when a model parameter space is large and cannot be explored in detail within a reasonable time. In particular, the lack of information about EEG feature importance can affect the results of Bayesian assessment of EEG maturity. In this paper we explore how the posterior information about EEG feature importance can be used to reduce a negative influence of disproportional sampling on the results of Bayesian assessment. We found that the MCMC integration tends to oversample the areas in which a model parameter space includes one or more features, the importance of which counted in terms of their posterior use is low. Using this finding, we proposed to cure the results of MCMC integration and then described the results of testing the proposed method on a set of sleep EEG recordings.
Keywords - EEG, brain maturity, Bayesian Model Averaging, Markov Chain Monte Carlo

1. Introduction

Early diagnosis of abnormal newborn brain development is a challenging problem in the developmental neurology and clinical neonatology. Experts attempt to assess brain maturity by visual analysis of age-related patterns in electroencephalograms (EEG) recorded from sleeping newborns (Tharp 1990; Holthausen et al. 2000). The analysis can take hours of expert work to confidently interpret sleep EEG, as the age-related patterns widely vary during sleep hours as well as between patients, and there are no regular rules for interpretation of these patterns (Cooper et al. 2003). There are neurological evidences that the post-conceptional ages (PCA) of healthy newborns normally match their EEG-estimated ages. In cases when the mismatch is observed during two and more weeks, the newborn's brain development is most likely abnormal (Scher 1997). Thus, the mismatch between PCA and EEG-estimated ages can alert about abnormal brain development.

In the first publications on EEG assessment of newborn brain development (Parmelee et al. 1968), the experts have visually analyzed 47 EEG recordings made in 11 PCA groups between 39 and 43 weeks. The experts have found 10 maturity-related EEG patterns. Then the experts have estimated the PCA of each EEG recording by counting the distribution of the maturity-related patterns. The expert estimates have exactly matched the stated PCA in 27.6% of cases. In 59.5% the matches were within ±1 week, and 77.5% of cases were found matching within ±2 weeks.

In later publications, it has been attempted to learn brain development models from sleep EEG data recorded from newborns whose maturation was preliminary estimated by experts. In (Scher, Steppe, & Banks 1996), the regression models have been applied to mapping the brain maturity into EEG index. In (Schetinin & Schult 2005; Crowell, Kapuniai & Jones 1978), the classification models have been used for distinguishing the maturity levels, at least, for one normal and one abnormal levels of brain development.

The above attempts were aimed at learning a single model providing the maximum likelihood on given EEG data. However such models cannot ensure the maximum accuracy when the likelihood distribution is affected by noise and its shape is multimodal. Besides, the model selection methodology cannot provide estimates of a full posterior distribution which is required for accurate assessment of the uncertainty in model outcomes.

In contrast, Bayesian classification enables the uncertainty to be accurately estimated via averaging over areas of high densities of the likelihood (Chipman, George, & McCullock 1998; Duda, Hart & Stork 2000; Denison et al. 2002; Armero et al. 2011). The estimates of uncertainty are made over an ensemble of classification models obtained during Bayesian averaging. The use of Decision Trees (DTs) as classification models enables to select features which make the most significant contribution to the classification. The feature selection becomes important when prior information on EEG feature importance is absent or deficient. Besides, DTs are attractive classification models as experts can interpret them. In the case of ensembles, a single DT providing a Maximum Posterior can be selected for interpretation as we proposed in (Schetinin et al. 2007).

The results of implementation of Bayesian averaging are critically dependent on the prior information and on the model parameter diversity in areas of averaging. When averaging is done over areas of interest with maximum likelihood, the resultant class posterior distribution is unbiased, and therefore the classification error is minimal. The use of prior information enables to specify the areas of interest and thus to improve diversity in model parameters.

Particularly, the prior information on EEG feature importance can be absent and so the areas of interest cannot be explicitly specified and then explored in detail (Domingos 2000; Schetinin & Maple 2002). Selection of EEG features has been shown improving the classification in (Yom-Tov & Inbar 2002).

In our previous work (Jakaite & Schetinin 2008), we attempted to mitigate the lack of prior information and proposed a new strategy for Bayesian averaging over DT models for predicting trauma survival. In this case of application, we observed that some screening tests (namely features) make a weak contribution to the model outcome and then we found that the DTs exploiting such weak tests can be discarded without affecting the accuracy of estimating the full class posterior distribution. In practice, it is important to reduce the number of features without an increase in the classification uncertainty, and the proposed method has been shown able to do achieve that.

The above findings motivated us to explore the discarding strategy in case of Bayesian assessment of newborn brain maturity from sleep EEG being represented by spectral power and statistical features. The importance of these features has not been explored yet in detail for a particular classification model such as DT. We will expect that the posterior information on EEG features will be effectively used within this strategy and the ensemble of DTs will be refined by discarding those models which exploit weak features. Similarly to the results obtained in our previous research, we will expect that the proposed strategy will reduce a portion of oversized DT models in the ensemble and the uncertainty in assessment will be decreased (Jakaite & Schetinin 2008; Jakaite, Schetinin, & Maple 2008).

Bayesian averaging over classification models is known as a theoretical methodology of achieving most accurate estimation of class posterior distribution. The estimate is calculated by integration of the posterior distribution over model parameters by using a stochastic integration known as Markov Chain Monte Carlo (MCMC) integration. The use of the Bayesian methodology allows experts to obtain the exhaustive information on uncertainty or risks in EEG assessment of newborn's brain. Therefore, the shape of the distribution becomes important for estimating the uncertainty in EEG assessment.

As part of this research, we will explore the shape of the class posterior distribution counted for a given PCA over DT models to answer the question whether a mismatch between the EEG estimate and PCA of the newborn causes a significant change in the shape. We assume that when PCA matches EEG estimate, the distribution shape tends to be symmetrical as the areas of interests are mainly located around one age category. On the contrary, for the mismatching cases the distribution becomes rather asymmetrical as the areas of interests are spread over different age categories. We will test our assumption on the EEG data to answer this question.

Overall, we expect to achieve the accuracy of the Bayesian assessment of brain maturity comparable to that obtained by experts. The accurate estimation of class posterior distribution provided by the Bayesian methodology will allow experts to obtain the exhaustive information on risk in EEG assessment of the newborn's brain maturity. The use of DT models which are transparent for users will allow EEG experts make new finding in the neurological assessment of newborn brain.

2. Problem Statement

Typically, EEG experts assess the newborn brain maturity in terms of PCA measured in weeks. Most experts agreed that the physiological ages of newborns are known in the range ±2 weeks post conception. The weeks of PCA are most often counted on the base of information obtained from a questionnaire of the mother. Ultrasound dating has been shown more accurate than that and normally undertaken on the first and second triple-months. The dates are typically replaced by the ultrasound estimates if the difference exceeds ±1 week in the first triple and ±2 weeks, in the second triple (Hoffman et al. 2008).

The newborn EEGs are typically recorded via the standard C3T3-C4T4 electrodes during a few sleeping hours. In our case, the EEG recordings have been made by the polysomnograph Alice 3 with a sampling rate 100 Hz. The raw data have been then processed with the Fast Fourier Transform over each 6-s epochs to be represented by the standard spectral power bands: Subdelta (0-1.5 Hz), Delta (1.5-3.5 Hz), Theta (3.5-7.5), Alpha (7.5-13.5), Beta 1 (13.5-19.5 Hz), and Beta 2 (19.5-25 Hz).

For our experiments, the EEG features have been made consisting of two groups, basis and extension ones. The features of the basis group represent the relative and absolute values of the above six spectral power bands calculated for the two electrodes and their sum, making them 36. The features of the extension group represent the non-stationarity of an EEG recording estimated with our technique as shown in (Jakaite, Schetinin, & Schult 2011). This technique estimates the distribution of the pseudo-stationary intervals in EEG. Using this technique, we made the extension group of features including the segment rate and 10 bins of the distribution histogram of the intervals ranging from 2-s to 20-s. These EEG features represent the information in the time domain. In particular, using the combined time and frequency EEG features has been shown improving the classification of EEG (Iscan, Dokur & Demiralp 2011). Finally, we added the ratio of slow-to-fast activities counted as the ratio of absolute spectral powers in Theta and Alpha bands, increasing the number of features in this group to 12.

The two feature groups together include 48 EEG features representing the EEG epochs. For our experiments, each EEG recording has been represented as a vector whose elements are the average values calculated over all epochs in the EEG recording. This is the typical way to represent each EEG recording as a vector in a multidimensional input space.

Note that although the above 48 features have been thought most informative for our experiments, we cannot state that there exists the prior information about the importance of either each feature or a feature combination considered within the given classification model. Therefore, using DT models for the Bayesian classification, we would explore the relative importance of the given EEG features and provide experts with the additional information about feature importance.

As mentioned in the Introduction, in the absence of the prior information, the results of Bayesian classification will likely suffer from disproportionally sampling the posterior distribution, as we cannot expect that a multidimensional model space will be explored in detail, and the areas of interest will be properly explored within a reasonable time.

Obviously, using DT models within the Bayesian methodology will give us a possibility to estimate the importance of EEG features in terms of frequency of their use. Having obtained this information, we could assume that if an EEG feature is rarely used in the DT ensemble, then this feature makes a weak contribution. Given a threshold we could find a set of such features and then could refine the ensemble by discarding those DTs which use these weak features. As discussed in (Jakaite, Schetinin, & Maple 2008), gradually increasing a threshold we can discard these DTs while their contribution to the ensemble outcome is negligible. As part of the research we will explore whether such a discarding technique is able to improve the accuracy of Bayesian assessment.

Note that a trivial strategy of using the posterior information is to remove the identified weak features from the data set and then rerun the Bayesian classification. Clearly, this will reduce a dimensionality of the problem and therefore a dimensionality of a model parameter space so that the areas of interests will be explored in more detail. However, the use of this strategy is limited by the computational power and time required for MCMC integration. Therefore, it will be interesting to compare the proposed discarding technique and this rerunning strategy in terms of the accuracy of assessment.

3. Bayesian Averaging over Decision Tree Models

For a DT given with parameters $\boldsymbol{\theta}$, the predictive distribution is written as an integral over the parameters θ:

$$p(y \mid \mathbf{x}, \mathbf{D}) = \int_{\theta} p(y \mid \mathbf{x}, \boldsymbol{\theta}, \mathbf{D})\, p(\boldsymbol{\theta} \mid \mathbf{D}) d\boldsymbol{\theta},$$

where y is the predicted class (1, …, C), $x = (x_1, \ldots, x_m)$ is the m-dimensional input vector , and $\mathbf{D}$ are the given labeled (or training) data.

This integral can be analytically calculated only in trivial cases. In practice, part of the integrand, the posterior density of $\boldsymbol{\theta}$ conditioned on the data $\mathbf{D}$, $p(\boldsymbol{\theta}|\mathbf{D})$, cannot usually be evaluated. However, using $\theta(1)$, …, $\theta(N)$ as the samples drawn from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$, we can write:

$$p(y \mid \mathbf{x}, \mathbf{D}) \approx \sum_{i=1}^{N} p(y \mid \mathbf{x}, \boldsymbol{\theta}^{(i)}, \mathbf{D}) p(\boldsymbol{\theta}^{(i)} \mid \mathbf{D}) = \frac{1}{N} \sum_{i=1}^{N} p(y \mid \mathbf{x}, \boldsymbol{\theta}^{(i)}, \mathbf{D}).$$

The above integral can be approximated by using a MCMC integration technique as described in (Chipman, George, & McCullock 1998; Denison et al. 2002). To perform such an approximation, we need to run a Markov Chain until it has converged to a stationary distribution. Then we can collect N random samples from the posterior $p(\boldsymbol{\theta}|\mathbf{D})$ to calculate the desired predictive posterior density.

Using DTs for the classification, we need to find the probability with which an input x is assigned by a terminal node to the jth class. The DT parameters are defined by s_i^{pos}, s_i^{var}, s_i^{rule}, $i = 1, \ldots, k-1$, where s_i^{pos}, s_i^{var}, and s_i^{rule} define the position, predictor and rule of each splitting node, respectively, and k is the number of terminal nodes. Having defined the parameters of a Markov Chain, we can specify the priors for the MCMC integration as follows. First, a maximal number of splitting nodes in DTs can be specified to be, for example, $s_{max} = n - 1$. Second we can draw any of the m attributes from a uniform discrete distribution $U(1, \ldots, m)$ and assign $s_i^{var} \in \{1, \ldots, m\}$. Ultimately, a candidate value for the splitting variable $x_j = s_i^{var}$ can be drawn from a discrete distribution $U(x_j(1), \ldots, x_j(L))$, where L is the number of possible splitting rules for variable x_j.

The use of these priors allows us to explore DTs which split data in as many ways as possible. Note that the DTs with different numbers of splitting nodes should be explored in the same proportions.

The number of splitting nodes can vary and therefore the MCMC has to integrate over a model parameter space of a variable dimensionality. For such integration, the MCMC technique is extended by Reversible Jump (RJ). The implementation of RJ MCMC integration over DT models has been proposed in (Chipman, George, & McCullock 1998; Denison et al. 2002) by using the following four moves:

Birth move randomly splits the data points falling in one of the DT terminal nodes by inserting a new splitting node with a variable and rule drawn from the given priors.

Death move randomly picks a DT splitting node with two terminal splits and then assigns it to be one terminal node with the united data points.

Change-split move randomly picks a splitting node and assign it to be with a new splitting variable and rule drawn from the given priors.

Change-rule move randomly picks a splitting node and assign it to be with a new rule drawn from the given prior.

We can see that the first two moves, birth and death, reversibly change the dimensionality of model parameter space. The third and fourth moves change the model parameters within a current dimensionality. Specifically, the change-split move enables to make “large” jumps which potentially increase the chance of sampling from areas of a maximum posterior whilst the change-rule move does “local” jumps.

The RJ MCMC technique starts drawing samples from a DT consisting of one splitting node with the parameters randomly assigned within the predefined priors. While a DT grows, its likelihood is increased and then tends to oscillate around a stable value. This phase is named burn-in and must be preset sufficiently long in order to achieve the stable posterior distribution. During the second phase

named post burn-in, the samples of the posterior distribution are collected for approximation of the desired class-posterior distribution.

4. Curing the Ensemble

As discussed in the Introduction, results of MCMC integration over DT models can suffer from disproportional sampling. The proposed method aims at refining an ensemble of DT models by discarding those models which use weak EEG features. The weak features are defined as those whose posterior probabilities of use in the DT ensemble do not exceed a given threshold. Clearly, we can count these probabilities when an ensemble of DT models has been collected during MCMC integration. The probabilities of using features are normally interpreted as feature importance within a given classification model.

Having obtained the information about feature importance, we need to set a threshold, a maximal probability, for which features are defined weak. Such a threshold can be easily found, if we sort out the features in an order of ascending importance.

To begin, the first feature with the least importance is selected as a candidate.

At the second stage, we find all DT models included in the ensemble which exploit this feature. Having found such DT models, we then label them as a set of candidates for discarding from the ensemble as weak models.

At the third stage, the set of candidate models is removed from the ensemble, and then its performance is tested on the given data.

Finally, at fourth stage, the proposed change is made accepted if its performance in terms of assessment accuracy and ensemble entropy remains within a given confidence interval (normally 2-sigma interval). If accepted, the next feature is selected from the list of ordered features, and the second stage is repeated. Otherwise, the algorithm ends up with the desired threshold.

Obviously, the larger the threshold, the greater number of features is defined as weak, and therefore the larger portion of DT models is discarded. In the following section, we describe the experiments with the proposed technique of refining a DT ensemble obtained within the Bayesian model averaging framework presented in the previous section.

5. Experiments

In our experiments, we used EEG data recorded from newborns during sleep hours in clinics. All the recordings have been assessed by EEG experts. The data were represented with a set of EEG features described in the above section 2. The newborns were assigned in 10 age groups or classes according to their weeks of PCA. We run these experiments first to test our assumptions described in sections 1 and 2 and second to test the proposed method described in section 4. The tests are made within the Bayesian methodology of averaging over DT models.

5.1 EEG Data and Settings

In our experiments we used 952 EEG recordings obtained in different clinics from newborns of age from 36 to 45 weeks of PCA as described in section 2. Each of the 10 age groups has been made including around 100 recordings. All the recordings were additionally tested on the presence of electrode move artifact causing a significant change or shift in EEG amplitude. The EEG recordings were also automatically tested on the presence of abnormal active or quiet sleep phases as well as of abnormal level of artifacts as we described in (Schetinin, Jakaite & Schult 2011).

The settings for running the Bayesian classification were made as follows. The number of DTs sampled in the burn-in phase was 100,000, and in a post burn-in phase 10,000. During the post burn-in phase each 10th model was collected in order to reduce the correlation between DT models. The minimal number of data samples allowed to be in DT nodes (or pruning factor) was set to five. Proposal variance was 1.0, and probabilities of making moves of birth, death, change-variable, and change-rule were set to 0.15, 0.15, 0.1, and 0.6, respectively. The performance and uncertainty of the DT ensemble collected in the post burn-in phase were evaluated within a 10-fold cross-validation.

Using the above settings for MCMC integration, we found that the rate of acceptance of DT models during the integration was around 0.23 in both phases. In the burn-in phase, the size of DTs was stabilized around 65 nodes after, on average, 10,000 samples, so that the remaining 90,000 samples were drawn from almost stable Markov Chain.

5.2 Importance of EEG features

As discussed in the Introduction, using DT models for classification within the Bayesian methodology allows us to count the importance of the EEG features in terms of the posterior probabilities of their use in DT ensemble. Fig. 1 shows the posterior probabilities of using all 48 EEG features in the basis (upper plot) and extension (lower plot) groups.

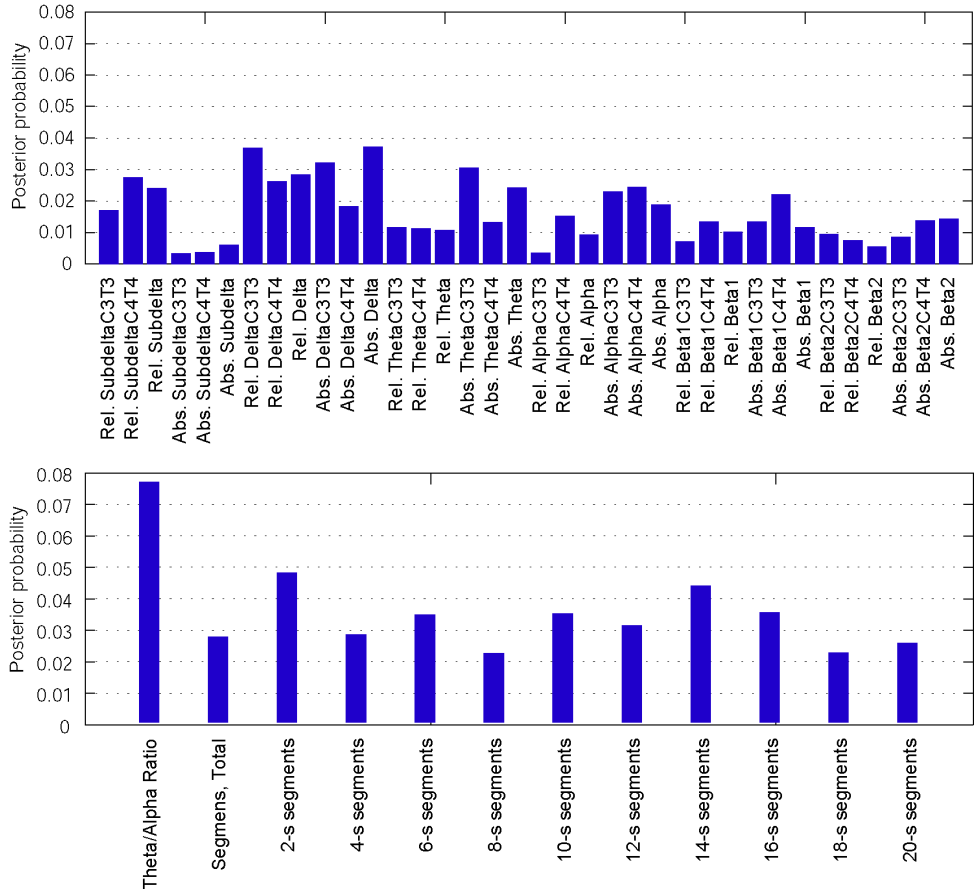

Fig. 1. Importance (posterior probabilities) of 48 EEG features characterising the relative and absolute spectral powers (upper plot) and the Theta-to-Alpha ratio and EEG non-stationarity (lower plot)

First, we see the importance of the features ranges between 0.0025 (AbsSubdeltaC3T3) and 0.078 (Theta/Alpha Ratio). Second, we observe that not all the features of the basis group are equally important, only 12 out of the 36 features are of the importance greater than 0.02. In contrast, the importance of all the features of the extension group is higher than that.

5.3 Refining the Ensemble

As discussed above, DTs collected in an ensemble use the EEG features in different combinations and thus the posterior probabilities of using them are different. The above Fig.1 shows that the probabilities or importance of the given features vary in a wide range. Some of the features with low importance are probably weak to make a distinguishable contribution to the classification. As

discussed in section 4, we can assume that discarding DTs using such weak features will improve the performance within the proposed method.

According to this method, we gradually increased the threshold T from 0 to 0.006 and then defined weak features. Table 1 shows the performance P, entropy E of the ensemble, and number k of EEG features found weak versus the threshold T; here k is the average over 10 folds. We can see that when the threshold $T = 0.003$, the number of weak features is $k = 8$, and when $T = 0.005$, k increases to 12.

Table 1. Performance *P*, entropy *E*, and number of weak features *k* versus threshold *T*

T	*k*	*P*, %	*E*
0.001	4	30.6±12.8	198.8±10.7
0.002	6	30.8±13.6	198.0±11.1
0.003	8	30.8±14.5	197.6±11.2
0.004	10	30.0±14.7	195.7±11.2
0.005	12	29.3±13.0	191.9±14.1
0.006	13	30.0±13.7	190.8±13.2

Fig. 2 shows the average performance and entropy obtained with the ensembles refined by discarding the 4, 6, 8, 10, 12, and 13 weak features, respectively. We see that the performance median slightly increases from 31% to 34% when the threshold is changed from 0 to 0.003. The uncertainty counted in terms of entropy of an ensemble is slightly decreased from 198.9 to 197.6. Further increasing the threshold to 0.006 leads to discarding 13 weak features without a significant drop in the performance. Clearly, the removal 13 out of 48 features makes the DT ensemble shorter and easier for interpretation; the expected number of rules must be decreased roughly on 1/3.

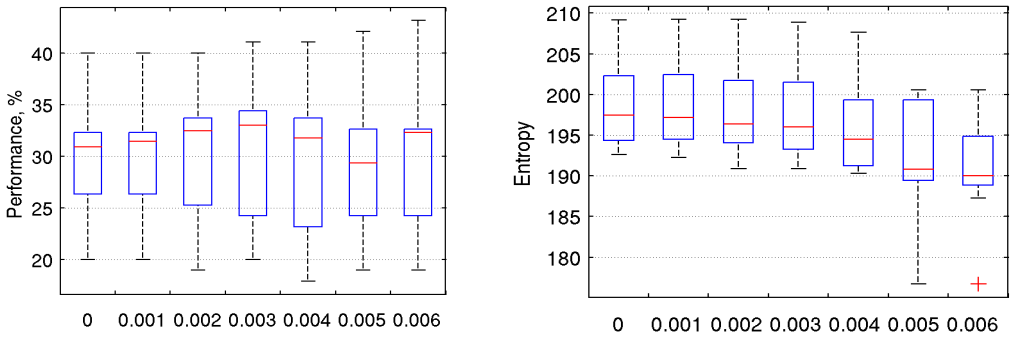

Fig. 2. Performance and Entropy over threshold *T*

We could assume that the rerun in the sense of retrain of the Bayesian classification on a data set without weak features will improve the performance. Indeed, the removal of weak features leads to reducing the dimensionality of the problem and subsequently to reducing the dimensionality of a model parameter space. However, when we reran the classification on a new data set without the eight weak features found at threshold of 0.003, we observed that the average performance was 27.1±6.9% and entropy 201.5±16.9. Because of extensive computations, we were limited to rerun the Bayesian classification on the other updated data sets. Nevertheless, we observe that the proposed technique outperformed the rerun strategy keeping the performance high and saving the computational time.

The box plots in Fig. 3 show the average numbers of DT splits for the original and refined ensembles as well as for the discarded DTs. The original ensemble included 10,000 DTs, but after

refining it included 5,800 DTs. We see that the median number of splits in the discarded set of DTs is 66.3, which is higher than that in the original ensemble. The median number of splits in the refined ensemble decreases to 65.5, as the portion of larger DTs has been removed. In the next subsections we will explore the accuracy of the refined ensemble of DTs.

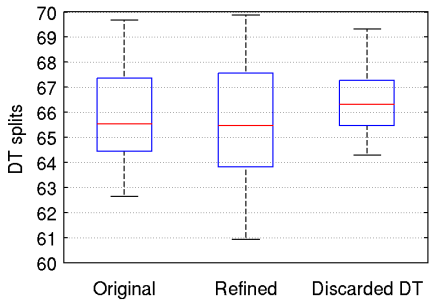

Fig. 3. Average number of DT splits in the ensembles

5.3 Extraction of Rules

It is important to note that DTs are hierarchical models which use features at different levels of the hierarchy, and therefore the use of a feature is defined by a chain from the DT root to a node testing the feature. In other words, the use of features in DT nodes has to be considered in the context of using the other features.

For illustration, let us show the fragments of the Maximum Posterior DT derived from the ensemble. This DT includes 70 nodes convertible into 71 probabilistic rules. Being limited in the space, we provide the seven rules for 36-week class (C=1) and the six rules for 45-week class (C=10) in the following notation:

C=1
N(1,1),N(2,0),N(3,1),N(4,0),N(5,0) -> 0.640
N(1,1),N(2,0),N(3,0),N(6,0) -> 0.455
N(1,1),N(2,1),N(7,1),N(8,0),N(9,0),N(10,0) -> 0.556
N(1,1),N(2,1),N(7,0),N(11,0),N(12,0),N(13,0) -> 0.571
N(1,1),N(2,0),N(3,1),N(4,0),N(5,1),N(14,1),N(15,0),N(16,1) -> 0.600
N(1,1),N(2,1),N(7,0),N(11,0),N(12,0),N(13,1) -> 0.250
N(1,1),N(2,0),N(3,0),N(6,1),N(17,1) -> 0.667
C=10
N(1,0),N(18,1),N(34,1),N(48,1),N(56,1),N(57,0),N(58,0),N(60,0) -> 0.500
N(1,0),N(18,1),N(34,0),N(35,0),N(39,1),N(63,0),N(70,0) -> 0.500
N(1,0),N(18,1),N(34,1),N(48,1),N(56,1),N(57,1),N(67,0),N(68,0) -> 0.727
N(1,0),N(18,1),N(34,1),N(48,1),N(56,1),N(57,0),N(58,1),N(59,1) -> 0.875
N(1,0),N(18,1),N(34,1),N(48,1),N(56,1),N(57,1),N(67,1) -> 0.562
N(1,0),N(18,1),N(34,1),N(48,0),N(49,1),N(66,1) -> 0.400

Here: N(1,*) = <38, 0.622>, N(2,*) = <41, 0.067>, N(3,*) = <22, 0.116>, N(4,*) = <16, 0.668>, N(5,*) = <43, 0.006>, N(6,*) = <33, 0.002>, N(7,*) = <23, 0.109>, N(8,*) = <22, 0.139>, N(9,*) = <37, 4.332>, N(10,*) = <3, 0.585>, N(11,*) = <2, 0.769>, N(12,*) = <14, 0.084>, N(13,*) = <40, 0.183>, N(14,*) = <20, 0.011>, N(15,*) = <41, 0.063>, N(16,*) = <26, 0.015>, N(17,*) = <27, 0.013>, N(18,*) = <43, 0.021>, N(34,*) = <45, 0.012>, N(35,*) = <23, 0.110>, N(38,*) = <29, 0.049>, N(39,*) = <45, 0.011>, N(48,*) = <16, 0.506>, N(56,*) = <44, 0.018>, N(57,*) = <45, 0.016>, N(58,*) = <11, 5.797>, N(59,*) = <13, 0.037>, N(60,*) = <37, 7.099>, N(63,*) = <3, 0.593>, N(66,*) = <10, 2.722>, N(67,*) = <16, 0.924>, N(68,*) = <1, 0.637>, N(69,*) = <41, 0.103>, N(70,*) = <37, 6.866>.

To illustrate the above rules, let us define $N(i,b) = \langle v,q \rangle$ as the ith node which takes threshold q to test feature v: if the value of the feature exceeds the threshold, then $b = 1$ and an input falls into the

right branch of the *i*th node; otherwise, $b = 0$ and the input falls into the left branch. For example, $N(1,*) = <38, 0.622>$ denotes the root node which compares variable #38 with threshold 0.622.

Let us also define a chain of nodes between the root and a terminal node as a rule providing a probability that a given input of the true class. For example, we can define a rule of class 1 as $N(1,1)$, $N(2,0)$, $N(3,1)$, $N(4,0)$, $N(5,0)$ -> 0.640, consisting of the five nodes which test variables in the following sequence <38, 0.622>, <41, 0.067>, <22, 0.116>, <16, 0.668>, <43, 0.006>. A given input finally falls into the left branch of 5th terminal node with probability 0.64. The value $p = 0.64$ of this probability is relatively high, as the alternative probabilities to be of a false class is roughly $(1 - p)/(1 - c) = 0.04$, where $c = 10$ is the number of classes.

Note that the above DT has been selected for the exact matching weeks. Obviously, this DT can be transformed to be used for EEG assessments within the other intervals ±1 and ±2 weeks.

5.4 Performance of EEG Assessment

Having obtained an ensemble of DT models, we calculated the performance of the Bayesian assessment within the 10-fold cross-validation. The calculation requires to count the number of newborns for which EEG estimates match their physiological ages of PCA. The matches were counted within the following three intervals: 0 week (exact match), ±1 weeks, and ±2 weeks. The ratio of these matches (that is the performance) within each of these intervals were 30.1% , 65.5%, and 85.1%, respectively. As discussed in section 2, the neurological assessment of newborn brain maturity is mainly made within ±2 weeks of PCA.

Table 2 shows the spread of age classifications over the given age groups from 36 to 45 weeks of PCA. The table columns present the numbers of classifications fallen into the age groups ranged from -6 to +7 weeks. Thus Column 0 shows the numbers of classifications fallen into the actual age groups (exact matches), Column -1 shows the number of classifications fallen into an age group which is less than the actual age group on one week, Column -2 -- less than on two weeks, etc.

Table 2. Spread of age classifications

	Mismatch (weeks)														
PCA	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	Total
36							46	24	11	5	3	2	0	1	92
37						32	33	16	10	5	3	1	0	0	100
38					18	29	18	15	5	3	2	1	1	0	92
39				8	9	10	15	17	13	5	3	2	1		83
40			4	5	6	11	26	16	14	6	2	2			92
41		1	5	2	11	15	26	15	10	3	3				91
42	0	2	2	6	16	12	18	17	15	12					100
43	1	1	3	10	9	19	14	19	23						99
44	0	0	3	14	12	10	35	29							103
45	2	2	1	3	5	25	62								100
Total	3	6	18	48	86	163	293	168	101	39	16	8	2	1	952

The Total column presents the number of EEG recordings in each age group. This column shows that the numbers of recordings in each group are similar. The Total row shows the numbers of age classifications fallen in the age groups ranged between -6 and +7.

Table 3 shows the performance of the expert assessment of EEG maturation described in (Parmelee et al. 1968) together with the performance of the Bayesian assessment calculated within the same five age groups from 39 to 43 weeks of PCA within the ranges ±1 and ±2 weeks. Note that these results have been obtained on different EEG recordings and different sample sizes: we used the 465 recordings (in the above five age groups), whilst the experts have assessed only 47 recordings.

Table 3. Performances of expert and Bayesian assessment within the two intervals

Interval, weeks	Expert, %	Bayesian classification, %
±1	59.5	53.7
±2	77.3	80.8

Such a difference in the sample sizes does not allow us to compare the results directly. Nevertheless, we observe that the Bayesian assessment within ±2 week interval, on average, slightly outperforms the expert assessment.

It is important to note that an EEG assessment obtained within the Bayesian methodology is provided with an accurate estimate of the uncertainty as we discussed in the Introduction. Below we describe our experiments and results in estimating the uncertainty for EEG assessment.

5.5 Estimation of Uncertainty

In this subsection we describe how the estimates of uncertainty obtained within the Bayesian assessment can assist experts to reduce possible errors. Having obtained an ensemble of DT models, first we calculate the desired estimates by using the original ensemble and then explore whether the estimates are improved by using the refined ensemble.

Second we explore the class posterior probabilities obtained within the Bayesian assessment for patients assigned in different age groups. The assignments can be made matching or mismatching the stated PCA. As discussed in sections 1 and 2, within our research we consider a mismatch as the case of abnormal newborn's brain maturity, and therefore it is important to identify risk of the mismatch by analysing the posterior probability.

In our experiments we used the ensemble of DTs obtained on the 857 cases to test the other 95 cases, roughly equally distributed over the 10 age groups of PCA. According to the spread of age classification given in subsection 5.4, a few EEG assessments were found mismatched the PCA within the ±2 week interval. Therefore we expect that the class posterior probability distribution obtained for a mismatched case differs from that obtained for a matched case. To justify this assumption, we selected two cases of 6th class (41 weeks of PCA), one matching and the other mismatching the stated newborn's PCA.

Fig. 4 and 5 show the class posterior probabilities for these cases. Here, the left side plots show the class posterior probability distribution over the 10 classes within the 1σ intervals computed over all the DTs included in the original ensemble.

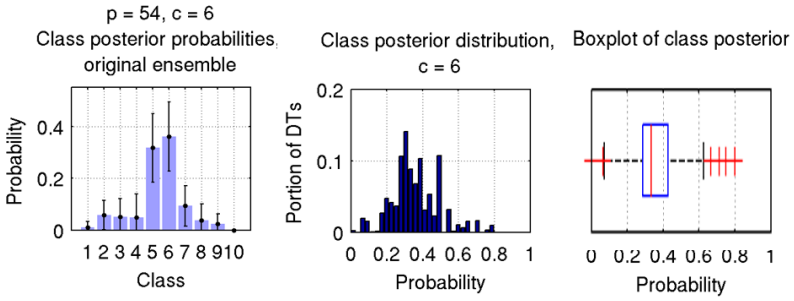

Fig. 4. Probability distributions estimated for matching case

From the left side plot of Fig. 4, we observe that the case belongs to the true class 6 with the average probability of 0.36, and with a slightly less probability this case belongs to class 5. Observing the intervals of these probabilities, we find that the EEG estimate matches the stated PCA within the acceptable interval of one week and thus we make the conclusion that the brain maturity of this newborn is normal.

The middle plot in Fig. 4 shows a distribution of all the DTs over probabilities that the case belongs to the true class 6. The average over this distribution gives us the maximal class posterior probability 0.36, observed in the left plot. The right side plot summarises the distribution and shows that its shape is rather symmetrical on the both sides of the median.

Let us now examine the probability distributions for the mismatching case shown in Fig. 5. In the left plot, we observe that this case can be wrongly assigned to the false class 1 with probability of 0.25. We also observe that the probabilities of classes 5, 6, 7 and 9 lie within the interval of class 1, and therefore we cannot make a confident conclusion on this case. The middle plot in this figure

shows a distribution of all the DTs over probabilities that the case belongs to class 6 stated for this newborn, and the right side boxplot shows that the shape of this distribution is clearly asymmetrical.

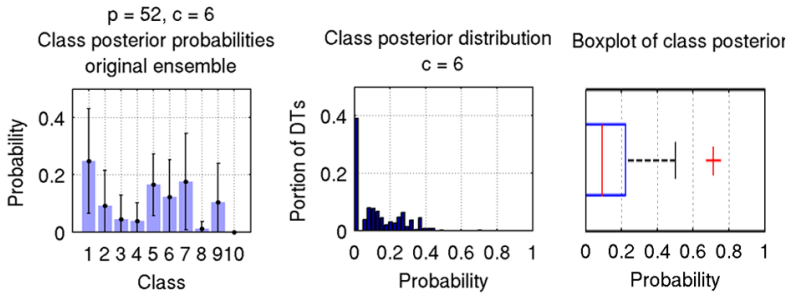

Fig. 5. Probability distributions estimated for mismatching case

The above two cases illustrate the use of the class posterior probability distributions obtained within the Bayesian methodology for estimating the uncertainty and reducing the risk of possible error. We observed that the distribution counted for the stated newborn's PCA (shown in the middle plots) becomes asymmetrical when an EEG assessment mismatches a newborn's PCA. The uncertainty can be quantitatively represented, and the shape asymmetry can be visually recognized.

As described in the previous subsection, the refined ensemble of DT models has improved the performance of EEG assessment, and therefore we can observe the corresponding changes in the class posterior distributions. Fig 6 shows these probabilities obtained with the refined ensemble for the above two cases.

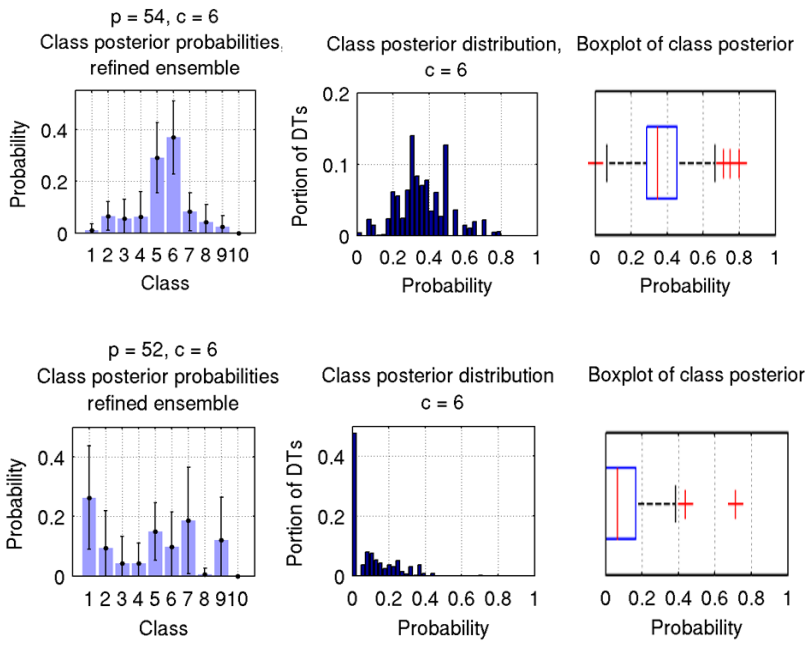

Fig. 6. Probability distributions estimated with the refined ensemble of DTs for matching (upper plot) and mismatching (lower plot) cases

The comparison shows that the average probabilities of classes 5 and 6 shown in the upper plot of Fig. 6 are slightly greater than those shown in Fig. 4 while their intervals are slightly smaller. As a result, using the refined ensemble decreases the uncertainty of EEG assessment. Comparing the posterior distribution obtained with the refined ensemble for the mismatching case shown in the lower plots of Fig. 6 and those obtained with the original ensemble (Fig. 5), we observe a similar decrease in the uncertainty of EEG assessment.

6. Conclusions

We explored how the posterior information about EEG features can be used for improving the results of assessment of newborns' brain maturity obtained within the methodology of Bayesian averaging over DT models. We assumed that the posterior information about feature importance can be used to find weak EEG features. According to this assumption, part of DT models included during MCMC integration in the ensemble use such weak features and thus do not make a significant contribution to the assessment.

We also observed that during the MCMC integration a candidate DT model, being assigned by chance to use a weak feature, can be accepted even with a decrease in its likelihood. In the presence of many weak features, such models will likely be disproportionally represented in the ensemble. Therefore discarding such DT models from the ensemble within the proposed technique could improve the EEG assessment.

The proposed technique has been tested on the EEG data recorded from newborns in the 10 groups of PCA. In our experiments the proposed technique has been shown capable of improving the proportions of DT models in the ensemble and as a result improving the performance of Bayesian assessment of newborn's brain maturity.

The results obtained with the proposed technique have been compared to those obtained by trivial rerunning the Bayesian assessment on a data set without weak features. We expected that the reduction of dimensionality of a model parameter space needed to be explored will improve the results. However the rerunning strategy has not been shown providing better performance that that provided with the proposed technique.

We also showed that an ensemble of DT models can be reasonably represented by a single DT providing the Maximum Posterior as a set of probabilistic rules transparent for experts. Each rule is formulated as a sequence of logical terms describing a chain between the DT root and one of the DT terminal nodes.

We expected to achieve the accuracy of the Bayesian assessment of brain maturity comparable to that obtained by experts. Although the EEG recordings used in our experiments were different, we found that the Bayesian assessments slightly outperform the expert assessments made in the same age groups.

We also expected to obtain the accurate estimation of class posterior distribution within the Bayesian assessment to provide experts with the exhaustive information on risk in EEG assessment of the newborn’s brain. Finally, in our experiments, we showed that the Bayesian assessment of the posterior probabilities are accurate and can be used for evaluating the risk of possible errors.

Acknowledgment

The authors are grateful to the Leverhulme Trust, UK, for funding this research.

References

Armero, C., Artacho, A., Lopez-Quilez, A., & Verdejo, F. (2011). A probabilistic expert system for predicting the risk of Legionella in evaporative installations. *Expert Systems with Applications*, 38(6), 6637-6643.

Chipman, H., George, E., & McCullock, R. (1998). Bayesian CART model search. *Journal of American Statistics*, 93, 935-960.

Cooper, R., Binnie, C., & Schaw, J. C. (2003). Neurophysiology of the neonatal period. In Binnie, C., Cooper, R., Mauguiere, F., Osselton, J. W., Prior, P. F., and Tedman, B. M. (Ed). Clinical neurophysiology: EEG, paediatric neurophysiology, special techniques and applications. Elsevier Science.

Crowell, D., Kapuniai, L, & Jones, R. (1978). Autoregressive Spectral Estimates of Newborn Brain Maturational Level: Classification and Validation. *Psychophysiology*, 15(3), 204-208.

Denison, D., Holmes, C., Malick, B., and Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Wiley.

Domingos, P. (2000). Bayesian Averaging of Classifiers and the Overfitting Problem. *Proceedings 17th International Conference. on Machine Learning*, San Francisco, CA., 223-230.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern Classification. Wiley Interscience. 2nd ed.

Holthausen, K., Breidbach, O., Scheidt, B., & Frenzel, J. (2000). Brain Dysmaturity Index for Automatic Detection of High-Risk Infants. *Pediatric neurology*, 22(3), 187-191.

Hoffman, C. S., Messer L. C., Mendola, P,. Savitz, D. A., Herring, A. H., & Hartmann, K. E. (2008). Comparison of gestational age at birth based on last menstrual period and ultrasound during the first trimester. *Paediatrics and Perinatal Epidemiology*, 22(6) 587–596.

Iscan, Z., Dokur,Z., Demiralp, T. (2011). Classification of electroencephalogram signals with combined time and frequency features. *Expert Systems with Applications*, 38(8), 10499-10505.

Jakaite, L., & Schetinin, V. (2008). Feature Selection for Bayesian Evaluation of Trauma Death Risk. *Proceedings 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*, 20, 123-126.

Jakaite, L., Schetinin, V., & Maple, C. (2008). Feature Importance in Bayesian Assessment of Newborn Brain Maturity from EEG. *Proceedings of the 9th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, edited by L. A. Zadeh et al, 191-195.

Jakaite, L., Schetinin, V., & Schult, J. (2011). Feature Extraction from Electroencephalograms for Bayesian Assessment of Newborn Brain Maturity. *Proceedings of the 24th International Symposium on Computer Based Medical Systems (CBMS-2011)*, Bristol.

Parmelee, A. H, Jr., Schulte, F. J., Akiyama, Y., Wenner, W. H., Schultz, M. A., & Stern, E. (1968). Maturation of EEG activity during sleep in premature infants. *Electroencephalography and Clinical Neurophysiology*, 24(4), 319–329.

Scher, M., Steppe, D., & Banks, D. (1996). Prediction of Lower Developmental Performances of Healthy Neonates by Neonatal EEG-Sleep Measures. *Pediatric Neurology*, 14(2), 137-44.

Scher, M. (1997). Neurophysiological Assessment of Brain Function and Maturation: a Measure of Brain Adaptation in High Risk Infants. *Pediatric Neurology*, 16(3), 191-198.

Schetinin, V., Schult, J. (2005). A Neural-Network Technique to Learn Concepts from Electroencephalograms. *Theory in Biosciences*, 124(1), 41-53.

Schetinin, V., Fieldsend, J.E., Partridge, D., Coats, T.J., Krzanowski, W.J., Everson, R.M., Bailey, T.C., & Hernandez, A. (2007). Confident Interpretation of Bayesian Decision Trees for Clinical Applications. *IEEE Transaction on Information Technology in Biomedicine*, 11(3), 312-319.

Schetinin, V., & Maple, C. (2007). A Bayesian Model Averaging Methodology for Detecting EEG Artifacts. Proceedings *15th International Conference on Digital Signal Processing*, 499-502.

Schetinin, V., Jakaite, L., & Schult, J. (2011). Informativeness of Sleep Cycle Features in Bayesian Assessment of Newborn Electroencephalographic Maturity, *Proceedings of the 24th International Symposium on Computer Based Medical Systems (CBMS-2011)*, Bristol.

Tharp, B. (1990). Electrophysiological Brain Maturation in Premature Infants: an Historical Perspective. *Clinical Neurophysiology*, 7(3), 302-314.

Yom-Tov, E., & Inbar, G. (2002). Feature Selection for the Classification of Movements from Single Movement-Related Potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10(3), 170-177.